High Performance Geospatial Modeling of Biospheric Resources

Jeff Pedelty & John Schnase
NASA Goddard Space Flight Center
Greenbelt, Maryland

25 June 2003

Invasive Species

A Top Environmental Issue of the 21st Century …

Economic Costs:

- *$137+ Billion / Yr*

  (Pimentel, et al. 1999; NISRC Management Plan, 2001)

Environmental Costs:

- *Decreased biodiversity, ecological services, etc.*

Human-Health Costs:

- *West Nile Virus, Malaria, etc.*

Agricultural Costs:

- *Crop pathogens, hoof-and-mouth, mad cow disease*

Notorious examples include:

Dutch elm disease, chestnut blight, and purple loosestrife in the northeast; kudzu, Brazilian peppertree, water hyacinth, nutria, and fire ants in the southeast; zebra mussels, leafy spurge, and Asian long-horn beetles in the Midwest; salt cedar, Russian olive, and Africanized bees in the southwest; yellow star thistle, European wild oats, oak wilt disease, Asian clams, and white pine blister rust in California; cheatgrass, various knapweeds and thistles in the Great Basin; whirling disease of salmonids in the northwest; hundreds of invasive species from microbes to mammals in Hawaii; and the brown tree snake in Guam.

As many as 50,000 now, hundreds new each year ...

Cheatgrass
(Bromis tectorum)

A classic example …

- *Arrived in N. America from Eurasia in late 19th Century*
- *By 1920s had spread throughout the US replacing native flora in over 40 millions hectares of the arid Intermountain West*

Biology

- *Self-fertile winter annual, establishes in the fall, flowers in early spring, large seed production, can occupy disturbed sites and all soil types, 20 cm roots with extensive lateralization, dies back in early summer …*

Cheatgrass
(Bromis tectorum)

Ecology *(A vicious cycle …)*

- *Early emergence reduces soil moisture thereby competing with native perennials for water*
- *Early die-back and impact on natives increases fuel load and fire frequency*
- *Re-establishes early in burned areas*

Impact on ecosystem processes

- *Changes hydrological cycles and biogeochemical cycles (esp. C, N)*
- *Alters fire regimes at landscape scales*
- *Changes landscape albedo, which affects energy budgets and climate*
- *Associated with other potentially unpredictable and cascading effects …*

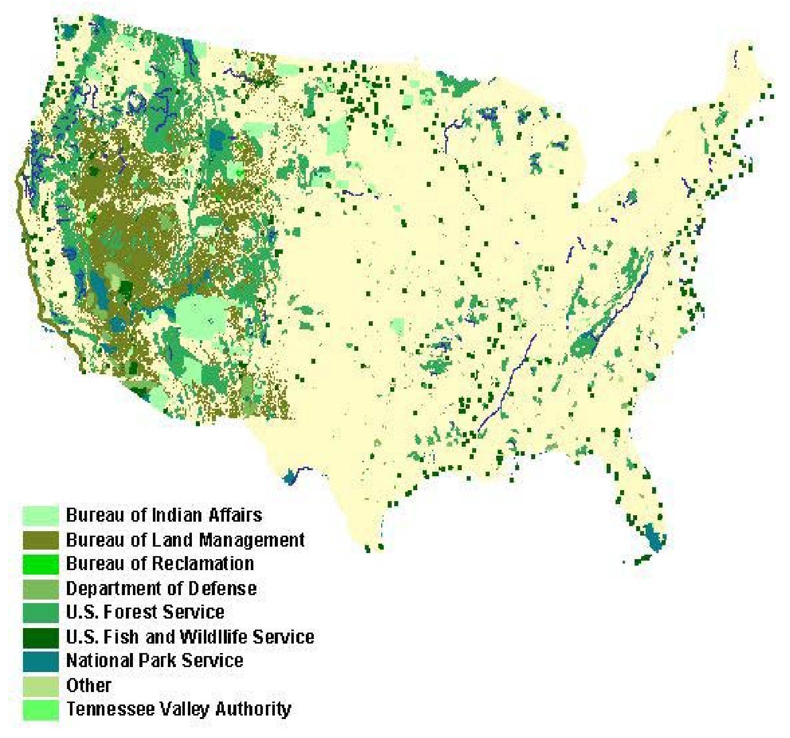
Federal Government Response

National Invasive Species Council (EO 13122 - 1999)

- *Co-Chaired by Departments of Agriculture, Commerce, and Interior*

<u>USGS</u> has a lead role in dealing with invasive species <u>science</u> in natural and semi-natural areas

- *Responsible for measurement, management, and control on all Department of Interior and adjacent lands ...*



USGS National Institute of Invasive Species Science

- USGS Biological Resources Division (BRD) laboratory
- Located at USGS's Ft. Collins Science Center
- New facilities opened Aug '02
- Director, Tom Stohlgren
- Many current / future partners ...

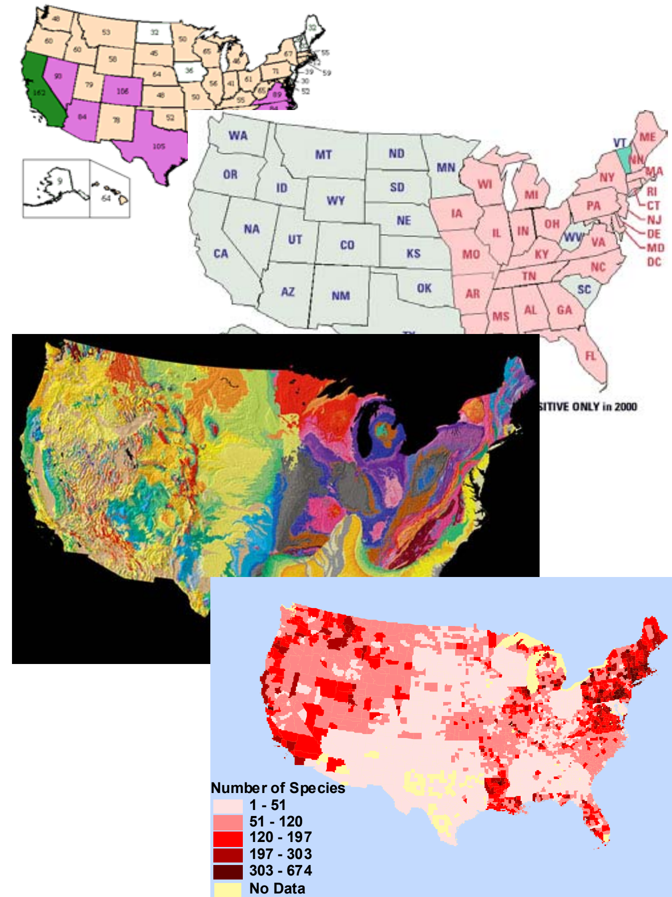
> **"Grand Challenge: Biodiversity and Ecosystem Functioning" with special emphasis on invasive species ...**
> ***NRC Committee on Grand Challenges in Environmental Sciences*, 2001**

> **"Needed: A National Center For Biological Invasions"**
> **Don Schmitz and Dan Simberloff**
> ***Issues in Science and Technology*, Summer 2001**

USGS Science / Client Needs

On-demand, predictive landscape- and regional-scale models and maps for biological invasions

- *Pick any point, land management unit, county, state, or region and determine the current invasion, and vulnerability to future invasion by species.*
- *Pick any species or group of species, and get current distributions, potential distributions, potential rates of change, and levels of uncertainty.*

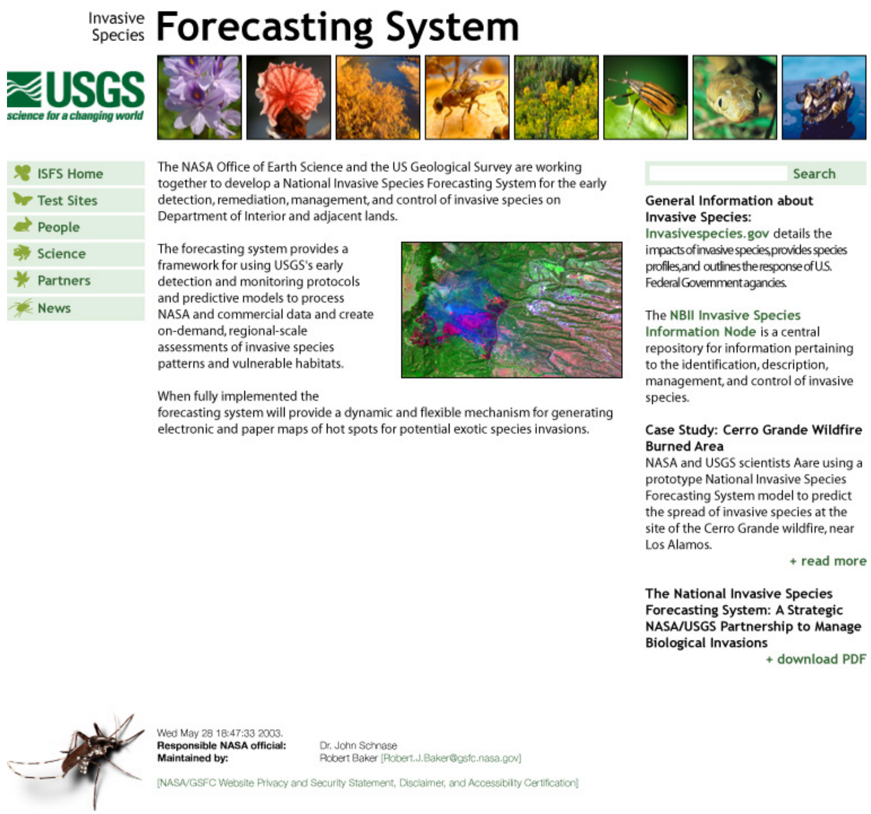


Data integration and sharing

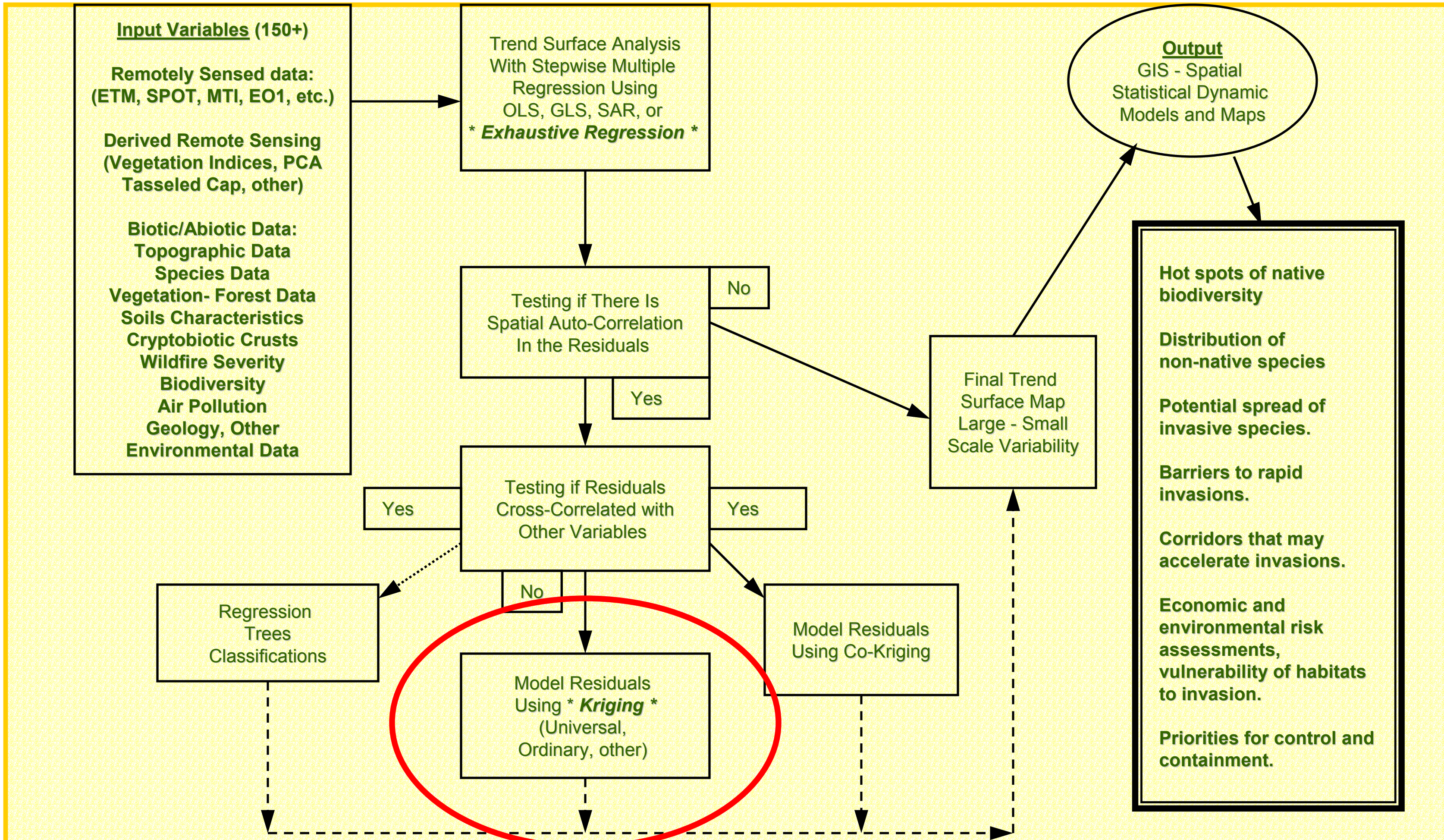
- *Comprehensive information on control efforts and cost. Share early detection data, control strategies, local expertise. Help public and private land managers.*

Public Interface

http://bp.gsfc.nasa.gov



USGS Predictive Modeling



What is Kriging?

Spatial interpolator

- A weighted linear combination of point measurements that exploits structure of spatial auto-correlation present in the data
- Spatial structure determines the appropriate weights for points that are “close” – allows for anisotropy
- Spatial structure is determined by modeling the empirical variogram – auto-correlation as a function of the separation distance
- Kriging determines weights by minimizing the variance of the errors: Best Linear Unbiased Estimator (BLUE)

<u>An Introduction to Applied Geostatistics</u>, Isaaks & Srivastava, 1989, Oxford University Press.

Why Kriging?

Stepwise regression is used to find the relationship between field samples and remote sensing, DEM, and ancillary data

Residuals (predictions from the stepwise regression minus observed value) are calculated for each sample point

Residuals are tested for spatial structure via viewing empirical variograms and statistical hypothesis testing (Moran's I)

If spatial structure exists Kriging is used to estimate the residual *surface* for the entire study area

Kriged residual surface is then added to the stepwise regression model to produce a final prediction that includes both small and large scale structure

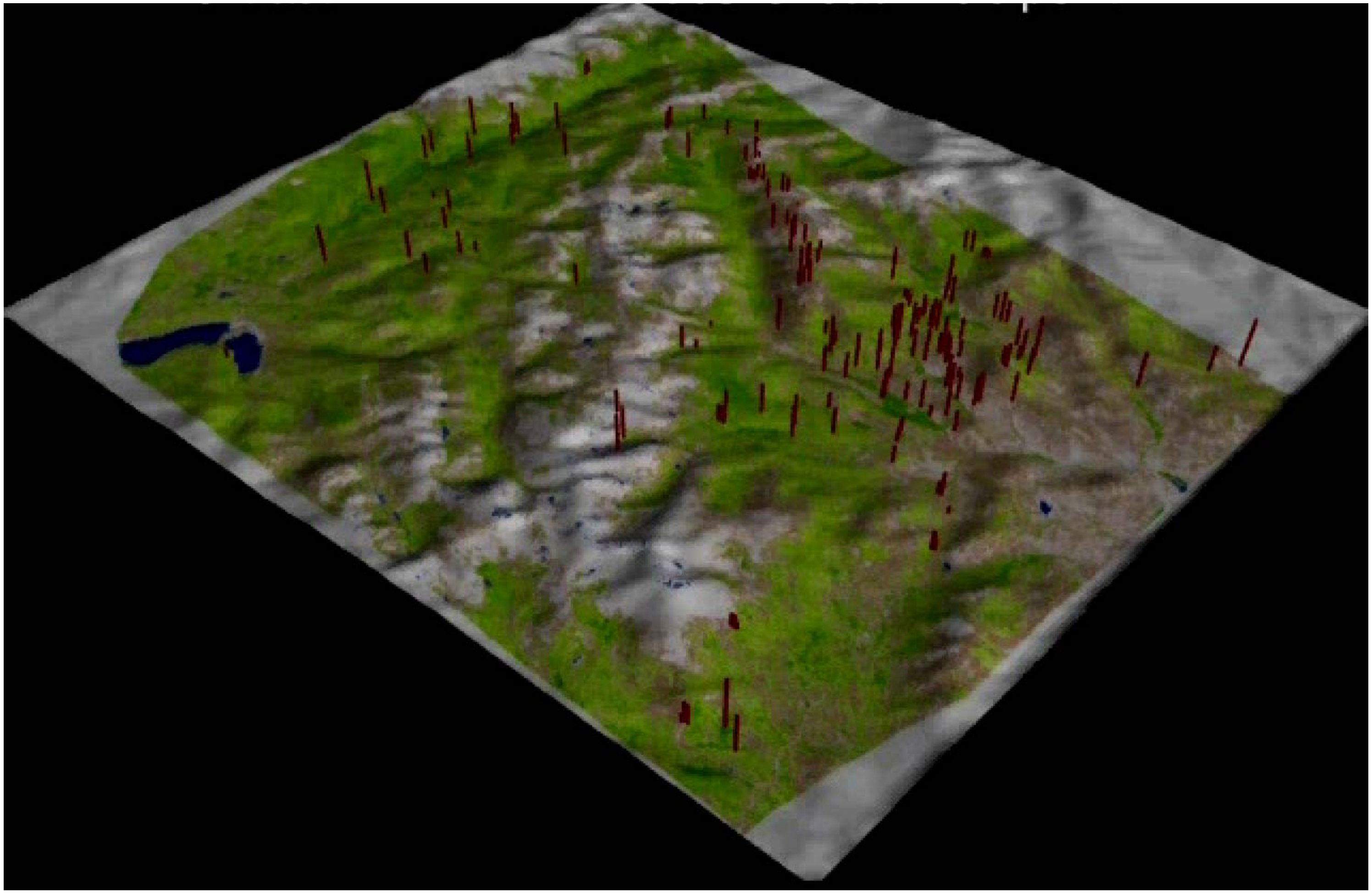
Why Parallel Kriging?

Kriging step in USGS processes has presented a major bottleneck.

Reducing the time of this computation allows different input variables to be considered, larger data sets to be incorporated, and more sites/locations to be modeled.

Kriging algorithms are widely used in general and parallel version has equally wide and general application.

Field Sampling in the Rocky Mountain National Park



Kriging Algorithm

Begin with ndata samples of quantity R

For each pixel in the output image

- Calculate distance from pixel to each sample data point (ndata x 1)
- Sort vector to find the nn nearest neighbor samples (nn x 1)
- Calculate covariance vector $\mathbf{D_j}$ for nearest neighbors (nn x 1)
- Calculate covariance matrix $\mathbf{C_{ij}}$ for nearest neighbors (nn x nn)
- Invert covariance matrix $\mathbf{C_{ij}}$
- Multiply by covariance vector to create weight vector (nn x 1)

$$\mathbf{W = C^{-1} \bullet D}$$

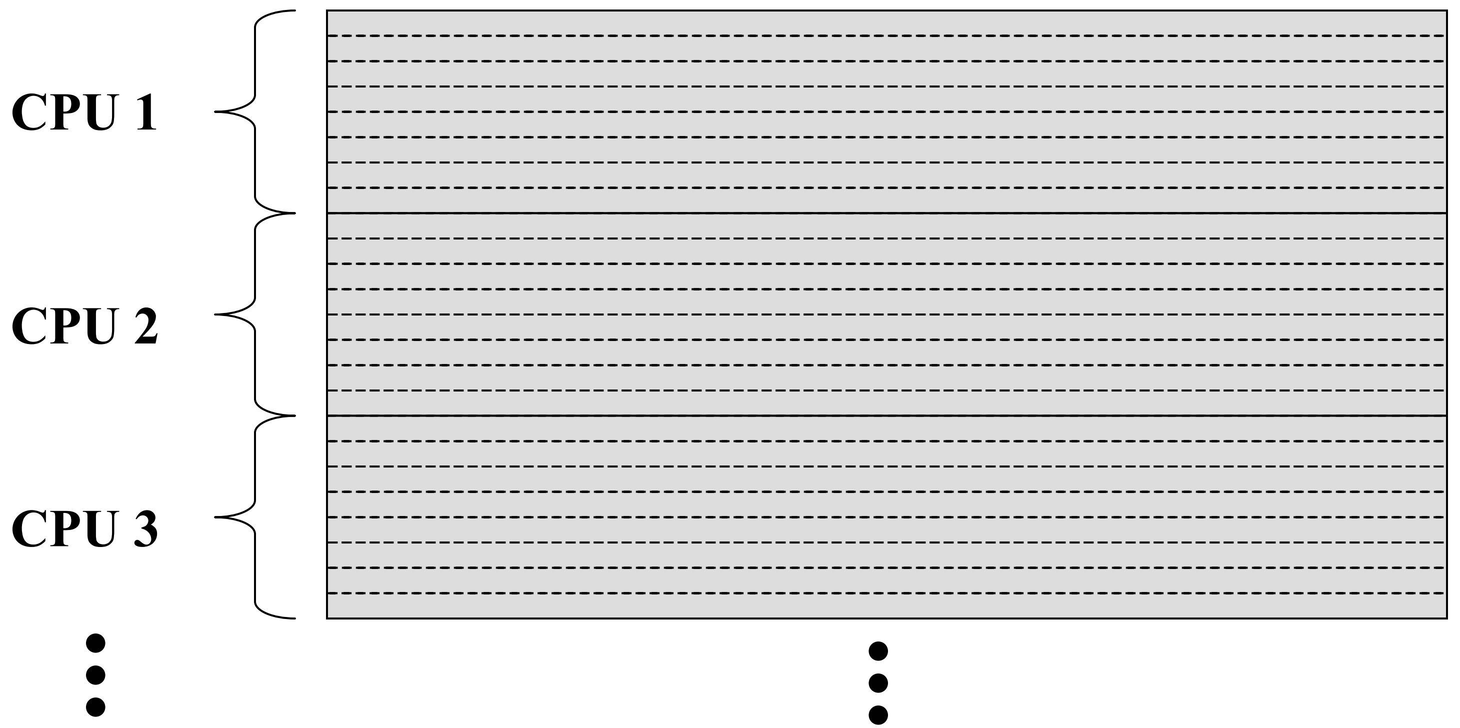
- Calculate dot product of data samples and weights to estimate R

$$\mathbf{R_{estimated} = W \bullet V}$$

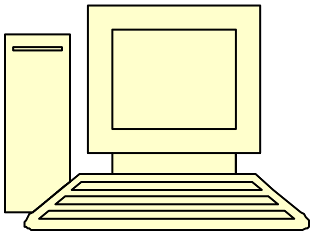
An Elegantly Parallel Algorithm

Parallelize using Domain Decomposition

Each processor gets a chunk of complete rows



Medusa / Frio Configuration



Frio on J. Schnase's desk (node 0)

Linux PC w/ 1.2GHz Athlon processor and 1.5GB memory

Gigabit Ethernet

Medusa Beowulf Cluster at NASA's GSFC

128-processor 1.2GHz Athlon MP

1GB memory on each dual-cpu node

2 Gbps Myrinet® internal network

MPI Implementation
Simple Version

Propagate input data from node 0 to all compute nodes

- X, Y sample locations and residuals at each point
- Desired size and resolution of output Kriged image
- Number of nearest neighbor samples to use
- Variogram information

Node 0 then starts MPI job on all compute nodes

Each compute node then:

- Determines its processor number and total number of CPUs
- Reads local data file
- Calculates its assigned rows
- Writes its rows (subimage) to local disk when finished

Node 0 grabs all files from each node

- Reassembles complete output image

MPI Implementation
Refined Version

Simple version does all computation before doing any communication

Refined version overlaps communication w/ computation

- At end of each row, each compute node issues asynchronous send (MPI_ISEND) of the row to node 0
- Processes next row while previous row is sent to node 0
- Issues wait/synchronize (MPI_WAIT) to verify receipt of previous row before sending current row.

Meanwhile, node 0:

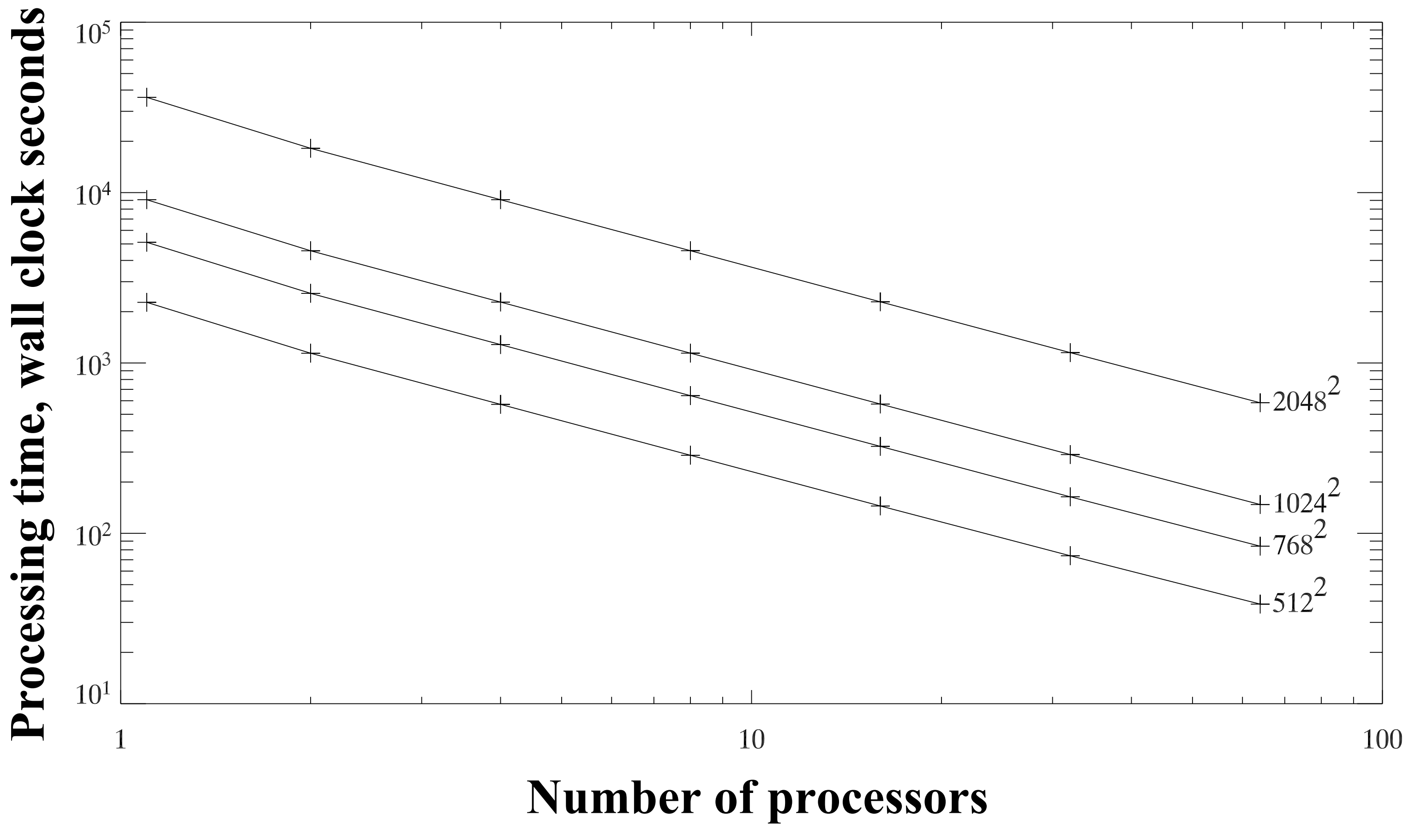
- Posts asynchronous receives from each compute node (MPI_IRECV)
- Issues MPI_WAITs to synchronize
- Builds output image row by row in memory
- Complete image is available when final compute node finishes

Timing Results

Wall-clock seconds on Medusa w/ Myrinet®

Number of Processors	Size of Kriged Image			
	2048^2	1024^2	768^2	512^2
64	583.9	147.5	84.0	38.4
32	1150.4	289.8	163.8	73.8
16	2285.1	573.89	324.0	144.7
8	4558.4	1142.0	642.8	287.2
4	9083.9	2277.4	1281.3	571.5
2	18190.4	4556.3	2562.0	1140.9
1	36252.7	9079.6	5107.0	2269.1

Scaling Curves



Scaling Results

Run time scales with area Kriged

- 2048^2 ran 16x longer than 512^2

Nearly linear scaling with # processors

Speedup Results

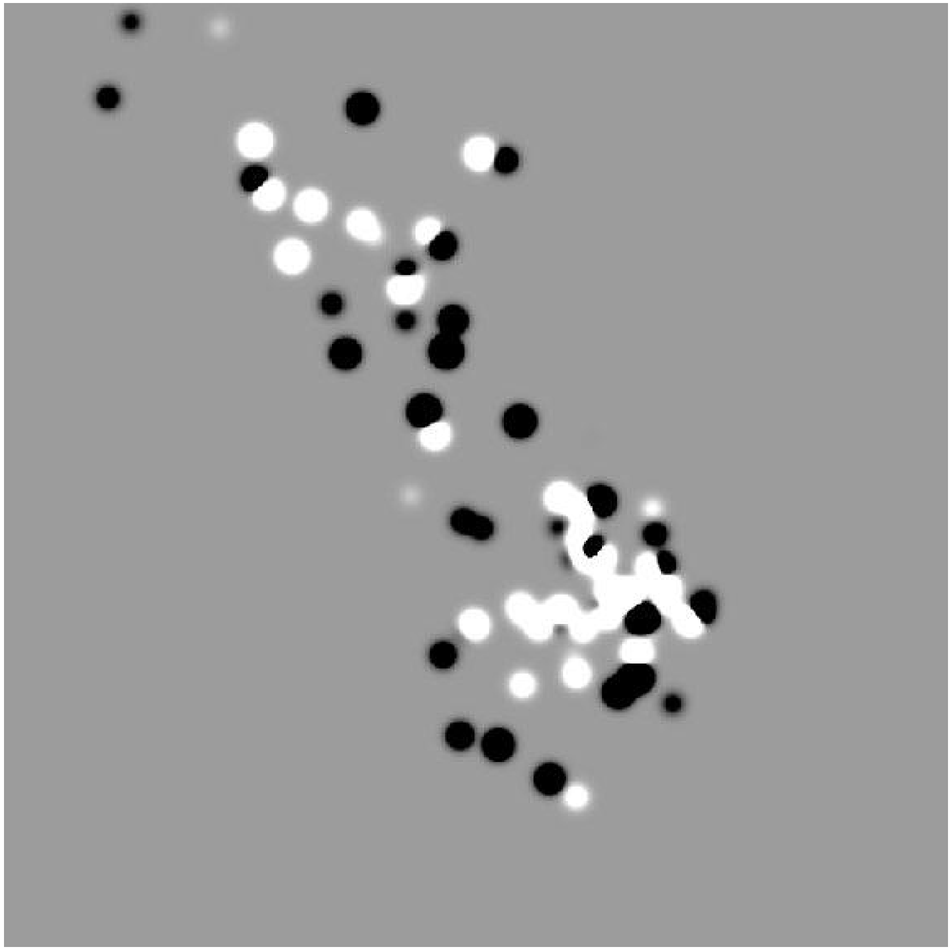
Number of Processors	Size of Kriged Image			
	2048^2	1024^2	768^2	512^2
64	62.1	61.6	60.8	59
32	31.2	30.9	30.5	29.7
16	15.6	15.4	15.3	14.9
8	7.8	7.7	7.7	7.5
4	3.9	3.9	3.9	3.8
2	1.97	1.96	1.95	1.92
1	1	1	1	1

Scaling Efficiencies

Number of Processors	Size of Kriged Image			
	2048^2	1024^2	768^2	512^2
64	97.0%	96.2%	95.0%	92.3%
32	98.5%	97.9%	97.4%	96.0%
16	99.2%	98.9%	98.5%	98.0%
8	99.4%	99.4%	99.3%	98.8%
4	99.8%	99.7%	99.6%	99.3%
2	99.6%	99.6%	99.7%	99.4%
1	100.0%	100.0%	100.0%	100.0%

The Kriged Residuals

Cerro Grande Fire Site



Future Work

CoKriging

- "multivariate Kriging" using spatial *cross*-correlation structure between residuals and input variables

Incorporation of new datasets

- MODIS time-series data
- NOAA precipitation data
- SRTM for Digital Elevation Models

Currently defining a large-scale problem

- Cheat grass on state or national level

Summary

Implemented MPI version of Kriging

- Nearly linear scaling
- ~31x speedup on 32 processor Beowulf cluster

Performance improvements allow expanded ISFS

- Test different models
- More data
- More sites